S-114.2510: Text Mining in Biological Databases

Santosh Tirunagari

January 25, 2011

Abstract

As the huge amount of the data growing in the Internet for the biological literature, It will a highly impossible task for managing the abreast of the developments. Hence there is a need of a technology which manages the overload. These methods are the Text mining techniques, which involve the processes of information retrieval, information extraction and datamining. By adding meaning to text, these techniques produce a more structured analysis of textual knowledge than simple word searches, and can provide powerful tools for the production and analysis of systems biology models.

Contents

List of Figures

0.1 Introduction

The text in the internet for biological datases is growing in large volumes since the start of the internet. There has been many techniques invented to manage, extract, identify, integrate ans extract the information from those texts. It is very important to extract the hidden information from the huge volumes of the text which are unknown.

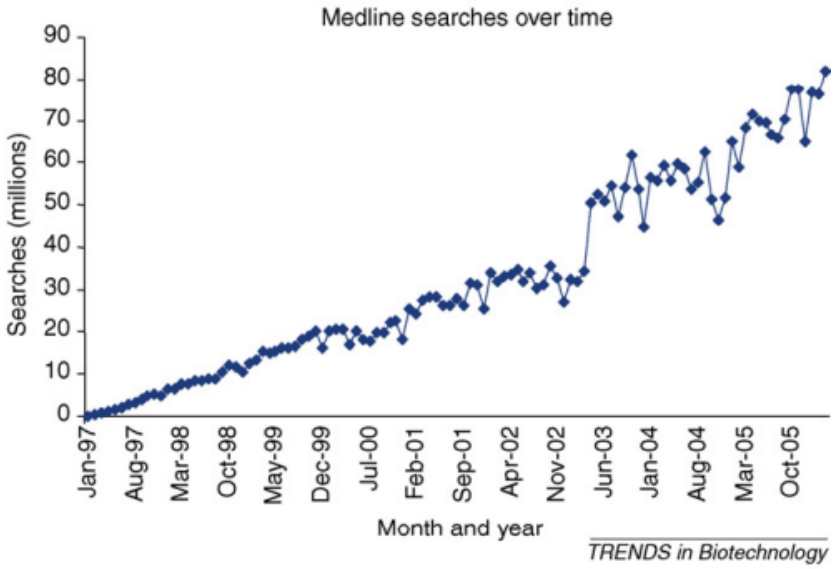

Figure 1: Increase in the biology literature from years [10]

The Figure 1 gives us the imformation of the growth in the biological text documents over the years. It is noteworthy to compare the number of MEDLINE1 searches in March 2006 (82.027 million) with the number in January 1997 (0.163 million). MEDLINE1 contains 15 million references to journal articles in the life sciences, and its size is increasing at a rate of more than 10% each year [10].

With the popularity of open access journal publishing, such as BioMed Central [1], full text articles are becoming more available. The availability of huge textual resources provides the scientist with the chance to search for correlations or associations such as protein–protein interactions [1,2] and gene–disease associations [3–6].

Text-mining (TM) in molecular biology is defined as the automatic extraction of information about genes, proteins and their functional relationships from text documents. TM has emerged as a hybrid discipline on the edges of the fields of information science, bioinformatics and computational linguistics. A range of text-mining applications have been developed recently that will improve access to knowledge for biologists and database annotators [10].

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data

[1] `http://www.biomedcentral.com/`

mining. These various stages of a text-mining process can be combined together into a single workflow [10].

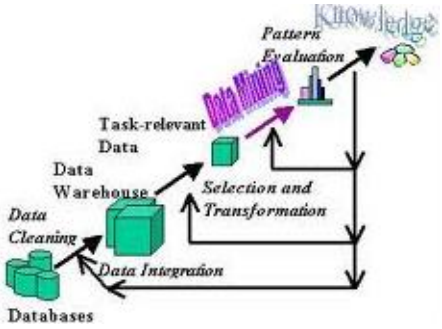

Figure 2: Text Data Mining

0.2 Finding relevant arcticles

The availability of huge textual resources provides the scientist with the chance to search for correlations or associations such as protein–protein interactions and gene disease associations.

is it difficult?

The context and the scenario of the same terms would be different. Hence it makes difficult to extract the information from the large text collection. for example we can consider the below examples.

- Proteins and genes are characterized within biological databases through unique identifiers. Each identifier is associated with its corresponding protein or nucleotide sequence and functional descriptions.
- Metabolites, proteins and genes often have a variety of names (terms) for denoting the same concept. For example, the metabolite glucose-6-phosphate is referred to as variants and permutations of a or b, D- or L-glucose (or hexose) -6- (mono) -phosphate.

Furthermore, within the same text a term can be given in an extended compounded form then later expressed through various mechanisms, including

- orthographic variation (usage of hyphens and slashes e.g. amino acid and amino-acid)
- lower and upper cases (NF-KB and NF-kb)
- spelling variations (tumour and tumor)
- Latin and Greek transliterations (oestrogen and estrogen)
- abbreviations (RAR and retinoic acid receptor)

0.3 Methods and Glossary

In this section, it is discussed about the evaluation methods, new measures and techniques for the quality of the information extracted is discussed.

0.3.1 Precision

precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$Precision = \frac{tp}{tp + fp}. \tag{1}$$

0.3.2 Recall

Recall is given by

$$Recall = \frac{tp}{tp + fn}. \tag{2}$$

0.3.3 Accuracy

The accuracy is the proportion of true results (both true positives and true negatives). It is a parameter of the test

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

0.3.4 F-measure

F-measure is the harmnic mean of Precision and Recall

$$F - measure = \frac{2.P \times R}{P + R} \tag{4}$$

0.3.5 SBML

The Systems Biology Markup Language (SBML) [2] is a representation format, based on XML, for communicating and storing computational models of biological processes. It is a free and open standard with widespread software support and a community of users and developers. SBML can represent many different classes of biological phenomena, including metabolic networks, cell-signaling pathways, regulatory networks, infectious diseases, and many others. It is the de facto standard for representing computational models in systems biology today.

0.3.6 bioPAX

BioPAX (Biological Pathway Exchange) [3] is a RDF/OWL-based standard language to represent biological pathways at the molecular and cellular level. Its major use is to facilitate the exchange of pathway data. Pathway data captures our understanding of biological processes, but its rapid growth necessitates development of databases and computational tools to aid interpretation.

[2] `http://en.wikipedia.org/wiki/SBML`
[3] `http://en.wikipedia.org/wiki/BioPAX`

BioPAX solves this problem by making pathway data substantially easier to collect, index, interpret and share. BioPAX can represent metabolic and signaling pathways, molecular and genetic interactions and gene regulation networks.

0.3.7 NER

Named entity recognition (NER)[4] (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

0.4 Solving difficulties

Ontologies are crucial for text mining because they provide semantic interpretation to text and also constrain the possible interpretations of biological entities (terms).When we provide semantic interpretation to text, we link terms to concepts in ontologies.

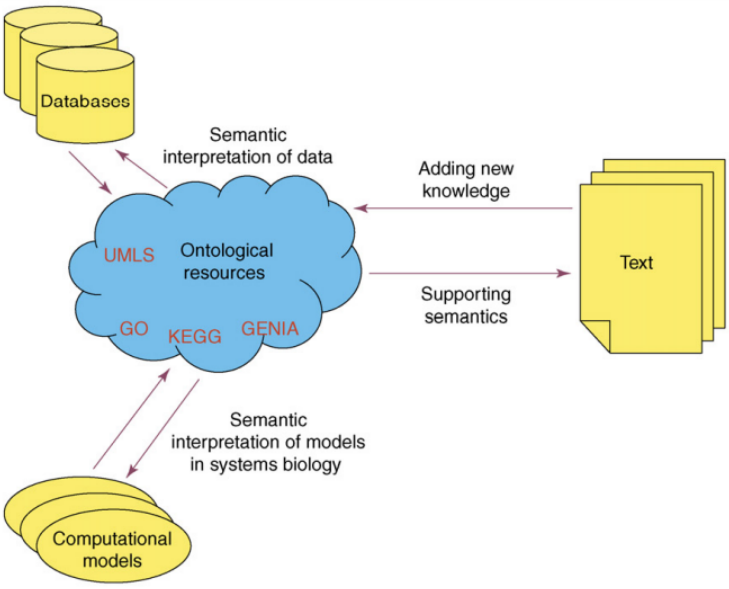

Figure 3: Importance of Ontologies [10]

Text mining tasks, particularly NER and IE, use the standard evaluation metrics of precision, recall and F-measure (see methods). Most text mining systems for NER and IE (relation extraction) use the F-measure to evaluate their results. The majority of systems compare their results with a 'gold standard': the most popular annotated corpora used by the text mining community as gold standards are GENIA [5] [11] and PennBioIE [6] [10].

[4] http://en.wikipedia.org/wiki/Named_entity_recognition
[5] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/
[6] http://bioie.ldc.upenn.edu/publications/latest_release/data/index.html

0.5 Term variation and Ambiguity

Term variation and term ambiguity make the identification of biological entities difficult.A particularly common term variation type in biology is representation by acronyms.

In MEDLINETM abstracts, 64 242 new acronyms were introduced in 2004, with the estimated total being 800 000 [8]. It was reported [9] that 5477 documents could be retrieved by using the acronym JNK, whereas only 3773 documents could be retrieved by using its full term, c-jun N-terminal kinase.

Acronym recognition aims to extract pairs of short forms. for example, ADM – adrenomedullin abductor digiti minimiadriamycin occurring in text. Existing methods for acronym recognition can be categorized into three groups:
1) heuristics and/or scoring rules
2) machine learning
3) statistical methods.

Term ambiguity occurs when the same term refers to many concepts. An example of term ambiguity is the term promoter, which refers to a binding site in a DNA chain, at which RNA polymerase binds to initiate transcription of mRNA by one or more nearby structural genes, whereas in chemistry it refers to a substance that in small amounts can increase the activity of a catalyst.

0.6 Information Extraction

Text is typically tokenized to identify the limits of words and sentences, then tagged (part-of-speech tagging) by assigning labels such as NOUN, VERB or ADJECTIVE to each word. Syntactic analysis identifies the basic textual chunks of a sentence, see Figure 5.

To detect and extract the types of evidence needed for hypothesis generation, we need semantic interpretation of the text, upon which we base relation extraction. Relation extraction extracts pairs or triples of biological entities, for example, p53 INDUCES Peg3 or Pw1 mRNA expression [10].

Few IE systems use deep linguistic knowledge. The advantage of full parsing [7] is that we can easily make generalizations for more than one type of biological interaction. To achieve this generalisation, we use predicate argument structures, which are canonical representations of sentence meanings that represent relations in an abstract manner, see Figure 6.

0.7 Potential applications in systems biology

Text mining techniques can be applied in a variety of areas of systems biology, and some applications are already beginning to emerge [12]. The structure, equations and parameters, including starting or fixed concentrations, define the ODE system and can be stored in a transmissible form as SBML (systems biology markup language [13]; A desirable goal now is to extend SBML to include the evidence for the models it describes. Another alternative is to have a BioPAX [14] file linked to a complementary SBML file (see methods and glossery).

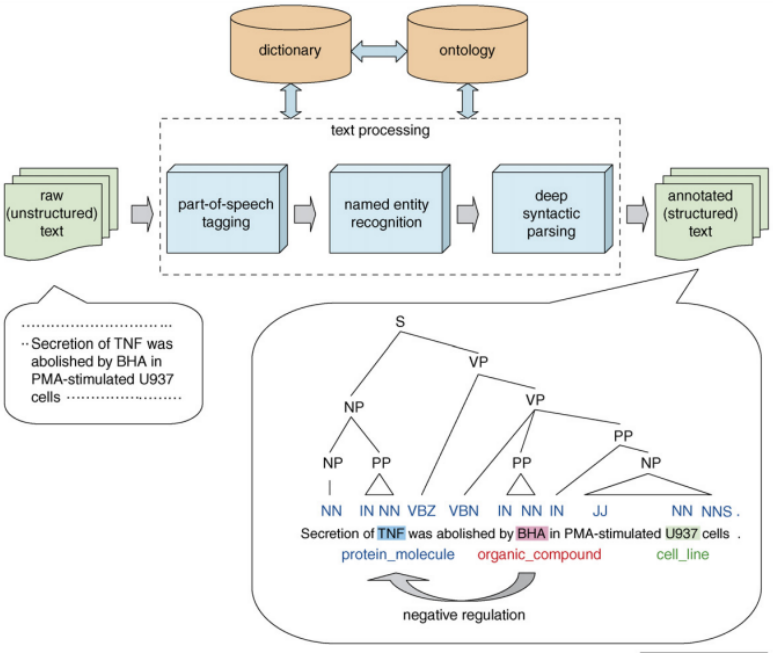

Figure 4: Text preprocessing [10]

0.8 Future directions

We consider that important future directions for the exploitation of TM in systems biology include the following.

- The availability of full texts is clearly of great significance because abstracts usually lack the relevant information. This is particularly true of the values of kinetic and binding parameters.
- A close integration of TM and DM techniques will benefit more widespread applications, for example, chemical structural similarity searches, the integration of medical records with genomic data and evidence from the literature for pharmaceutical applications. This will best be done in a distributed manner.
- Visualization from text mining results. Current visualization methods are still rather crude and there is much room for improvement here (NeRV at our ICS Lab).
- Better benchmarks for evaluating text mining tools that are relevant to biological needs.

0.9 Conclusion

Although the exploitation of text mining technologies is still in its early phases, they are now becoming sufficiently mature that they can be expected to become tools in the armory of every biologist and biotechnologist.

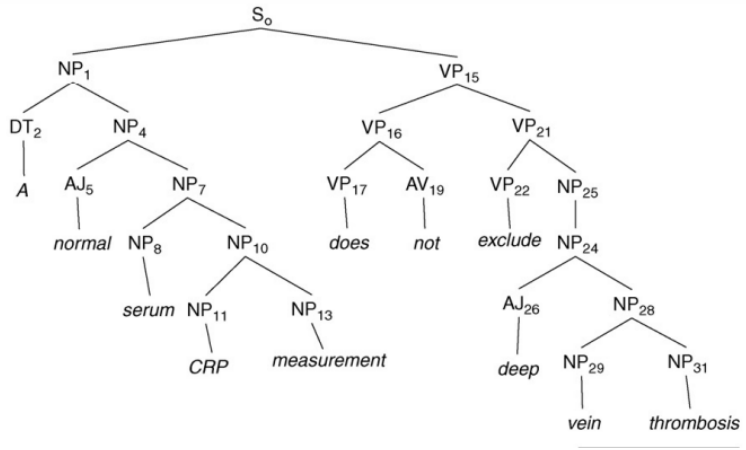

Figure 5: Parts of speech Tagging [10]

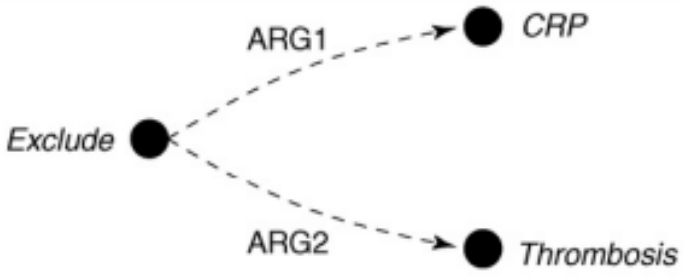

Figure 6: A predicate argument structure (PAS) [10]

Bibliography

[1] Ono, T. et al. (2001) Automated extraction of information on protein–protein interactions from the biological literature. Bioinformatics 12, 155–161

[2] Blaschke, C. et al. (2002) Information extraction in molecular biology. Brief. Bioinform. 3, 154–165

[3] Hao, Y. et al. (2005) Discovering patterns to extract protein–protein interactions from the literature. Part II. Bioinformatics 21, 3294–3300

[4] Korbel, J. et al. (2005) Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol. 3, e134

[5] Marcotte, E.M. et al. (2001) Mining literature for protein–protein interactions. Bioinformatics 17, 359–363

[6] Chun, H.W. et al. (2006) Extraction of gene disease relations from MEDLINE using domain dictionaries and machine learning. In Pacific Symposium on Biocomputing 2006 (Altman, A.B. et al., eds), pp. 4–15, World Scientific Publishing Co.

[7] Huang, M. et al. (2004) Discovering patterns to extract protein–protein interactions from full texts. Bioinformatics 20, 3604–3612

[8] 57 Chang, J.T. and Schutze, H. (2006) Abbreviations in biomedical text. In Text Mining for Biology and Biomedicine (Ananiadou, S. and McNaught, J., eds), pp. 138–165, ARTECH House

[9] Wren, J.D. et al. (2005) Biomedical term mapping databases. Nucleic Acids Res. 33, D289–D293

[10] Sophia Ananiadou, Douglas B. Kell and Jun-ichi Tsujii, Text mining and its potential applications in systems biology.

[11] Kim, J. et al. (2003) GENIA corpus – a semantically annotated corpus for bio-text mining. Bioinformatics 19 (Suppl. 1), i180–i182

[12] Miyao, Y. et al. (2006) Semantic retrieval for the accurate identification of relational concepts in massive textbases. In Proceedings of Coling/ACL Conference 2006, pp. 1017–1024

[13] Hucka, M. et al. (2003) The systems biology markup language (SBML), a medium for representation and exchange of biochemical network models. Bioinformatics 19, 524–531

[14] Luciano, J.S. (2005) PAX of mind for pathway researchers. Drug Discov. Today 10, 937–942